Received: March 17, 2025. Revised: May 20, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Identify Vulnerability of Adversarial Attack on Medical Images Using Deep Residual Shrinkage Network with Total Variation Loss Function

Amudha Gopalakrishnan[1]* Nalini Joseph[1] Umarani Srikanth[2]

[1]*Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, India*
[2]*Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India*
* Corresponding author's Email: amudhag.cse@gmail.com

Abstract: In medical image analysis, adversarial attacks introduce subtle distortions in medical images that makes deep learning models analyze the disease, which leads to inaccurate results. However, several various deep neural networks and transformer-based models used to identify the attack in medical images, which have failed to detect these attacks accurately due to noise perturbations and redundant features. To overcome this limitation, a Deep Residual Shrinkage Network with Total Variation Loss (DRSN-TVL) function is proposed to identify adversarial attacks and classify medical images accurately. The proposed TVL function is incorporated into the DRSN model, which minimizes unnecessary variations in pixel intensities that ensures the perturbations in the image remain smooth and realistic. Initially, the medical images are acquired from the Chest X-Ray and Eye disease datasets. Next, CLAHE is used to enhance the image contrast, and then a Wasserstein Generative Adversarial Network (WGAN) is used for generating fake images to integrate the adversarial attacks. After that, an EfficientNet-B7 model is employed for efficient feature extraction, and finally, identification of attacks is performed by the proposed model. The experimental results of DRSN-TVL method achieved an accuracy of 98.45%, which is higher than existing method ReseNet-153 V2.

Keywords: Adversarial attacks, Deep residual shrinkage network, EfficientNet-B7, Medical image analysis, Total variation loss function.

1. Introduction

Image analysis is among the essential parts of Artificial Intelligence (AI) technologies. It plays a significant role in various healthcare and medical fields, including radiology, pathology, and ophthalmology. Recently, the development of Deep Learning (DL) algorithms has made significant progress in their ability to contribute to medical diagnoses [1, 2]. The DL models show great potential in their ability to detect and diagnose diseases accurately that range across different domains that include genetic, infectious, deficiency, autoimmune, degenerative, and mental diseases [3]. Convolutional Neural Networks (CNNs) are extensively utilized in various applications involving image retrieval, image classification, and object detection [3, 4]. Also, Deep neural networks (DNNs) are increasingly being utilized in various fields and have become particularly crucial in safety-critical applications like malware detection, medical imaging, and so on [5] [6]. Despite their capabilities, DL models are often vulnerable to external Adversarial Attacks (AA) because these models heavily depend on the data on which they are primarily trained [7].

AA refer to images that are affected by slight perturbations, which means inserting a distortion that looks like the actual image [8]. These adversarial images are purposely generated to fool the DL-based prediction or classification models to provide inaccurate results. AAs in medical images use small perturbations that produce a misclassification, which are imperceptible to human vision [9]. Several AAs have been proposed based on having detailed information available and access to the target model, comprising hyperparameters and gradients [10]. The attacks that use model information are named white-

box attacks. Some of the AAs considered in medical images are Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD, Noise perturbation, and various Auto Attack methods [11-13]. Recently, transformer-based approaches like Vision Transformers (ViTs) are utilized in existing research for adversarial training along with the Generative Adversarial Network (GAN), which attained better adversarial attack identification [13] [15]. However, when it comes to the robustness of the ViTs model against AA, it is limited in identifying multiple attacks and fails to identify the attacks accurately. To overcome these limitations, a Deep Residual Shrinkage Network with Total Variation Loss (DRSN-TVL)-based classification model is proposed to identify and categorize AAs in medical images efficiently. The key contributions of this research are given as follows:

- In preprocessing, a Wasserstein GAN (WGAN) is utilized to enhance convergence and enhance adversarial attack classification by generating synthetic medical images with AA.
- A pre-trained model, namely EfficientNet-B7, is used to extract features efficiently by a compound function from both actual images and generated attack-based medical images.
- The proposed DRSN model utilizes soft-thresholding within the residual block that enables the model to not focus on irrelevant features, which helps to classify actual images and adversarial images efficiently. Also, the Total Variation Loss (TVL) helps the DRSN to stop learning sharp noise and makes it easier to identify unusual, unnatural patterns that indicate an attack.

This research paper is organized as follows: Section 2 discusses review of existing research. Section 3 describes proposed methodology. Results and discussion of the proposed method are presented in Section 4. The conclusion of this research paper is given in Section 5.

2. Literature survey

Vasan and Hammoudeh [16] developed a classification model to enhance resilience against AA based on advanced feature transformation training. The developed feature transformation using a fine-tuned ResNet152V2 network, was trained on actual medical images that helped the model adapt to medical imaging features effectively. The developed transfer learning-based adversarial training model utilized both original and adversarial images to enhance the robustness of the model against AA.

Sheikh [17] suggested a model to diagnose adversarial threat in CT image models based on dual-stage inference-time defense. Total Variation Minimization (TVM) and Non-Local means (NL-means) techniques were used to reduce noise in the medical images and enhance the quality of the images, which is especially useful for denoising complex medical images. Fuzzy Image Transformations (FIT) were used to enhance the robustness of the model by modifying the image features. DenseNet-121 was used to extract relevant features effectively and also used for classification, which helps improve the reliability of DL models in medical image classification, especially in the presence of AA.

Laykaviriyakul and Phaisangittisagul [18] introduced a CNN-based adversarial attack identification model to protect from AA. The introduced attack model utilized a perturbation noise filter for eliminating the noise from the adversarial samples before prediction, and a DiscoGAN framework was used to learn the relationships between original and adversarial images, which extracted the most relevant features effectively and improved the robustness of DL models against AA. The CNN was used to classify the reconstructed, denoised images and improve the accuracy.

Manzari [19] developed a method for generalizing medical images based on a robust vision transformer. A data augmentation technique was used to do random rotations, zooming, flips, and cropping of the images to simulate different variations in the data and make the model more robust. CNN was utilized to extract local features from images, and transformers were used to capture long-range dependencies. The hybrid CNN-transformer was used for classification and combined both local and global information for accurate predictions. The hybrid approach helps the model be more efficient and robust against AA in medical image classification.

Dai [20] developed a method to improve the adversarial robustness of medical image systems based on global attention noise. A noise injection method was used to add noise in the stages, which helps the model be more robust to adversarial perturbations. This method strengthens essential features and weakens generalization features, and the added noise makes the model smooth. Global Attention Noise (GATN) was used to improve the feature extraction process, which helps the model focus on important features and smooth decision boundaries. The Deep Neural Network (DNN) was used to classify the medical images effectively and improve robust accuracy.

Alzubaidi [21] suggested an Ensemble Feature Fusion (EFF) model based on deep learning and a machine learning based classification model for adversarial attack identification. The suggested EFF model utilized multiple XcepitonNet, which are pre-trained model for feature extraction that efficiently extracts relevant features. Also, machine learning models such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) were used for classification process. The main advantage of the suggested EFF model was its ability to incorporate different AAs without requiring training from scratch. However, the extracted features from the EFF model were fuses and directly used for classification, which consisted of irrelevant features that led to inaccurate classification results.

From the literature survey, it is observed that existing detection and classification models have limitations in identifying AAs in medical images. Limitations such as irrelevant features, poor image quality, and difficulty in differentiating actual and perturbation presented images due to subtle noise, which leads to inaccurate results. To overcome these limitations, a DRSN with a TVL function is proposed to enhance the identification of AAs in medical images by learning the difference between actual and adversarial images. A GAN-based model is utilized to generate the images with Aas, and the CLAHE technique is used to enhance image quality. For efficient feature extraction, EfficientNet-B7, a pre-trained model is used in this research to capture important features from the medical images.

3. Methodology

The objective of this research is to identify the AA and categorize the medical images using the proposed DRSN-TVL, which includes six phases: Dataset acquisition, preprocessing, Generative Adversarial Network, AA, Feature extraction, and proposed classification. Fig. 1 represents a block diagram of the proposed identification of the AA and classification of the medical images.

3.1 Dataset acquisition

To train a DL-based classification model for identifying the AA in medical images, four benchmark datasets are utilized in this research: a Chest X-ray dataset, Fundoscopy dataset, 3D – CT Lung nodule dataset, and Eye datasets, respectively.

3.1.1. Chest X-Ray dataset

Chest X-ray dataset [22] is one of the publicly available datasets that is widely used to detect and classify pneumonia disease. This dataset is an extended version of the Chest X-ray8 dataset, which comprises six additional thoracic diseases, namely, “edema”, “pneumothorax”, “pneumonia”, “fibrosis,” and so on. The Chest X-Ray dataset consists of 112,120 X-ray images, which are the frontal views obtained from 30,805 distinct patients. Every X-ray image in this dataset is multi-labeled with 14 various thoracic diseases.

3.1.2. Fundoscopy

This dataset [23] is commonly used for diabetic retinopathy detection and classification tasks. It comprises 3662 images, which are categorized into five classes based on the retinopathy states. The different stages of Diabetic Retinopathy (DR) are: Mild, Moderate, Severe, Proliferative DR, and Normal or No DR, respectively. These medical images are fed as input to the GAN model to generate synthetic images.

3.1.3. 3D-CT Lung Nodule dataset

This dataset [24] comprises 1018 Computed Tomography (CT) scan images, which is a publicly available dataset annotated by four specialist radiologists. From this dataset, 1670 nodules are randomly extracted with 33 slices. These data are labeled based on their malignancy scores into benign and malignant nodules, which are provided by the radiologists.

3.1.4. Eye dataset

This dataset [25] is an open access dataset, which comprises retinal images used in this research. This dataset is divided into four classes, namely Normal, Cataract, Glaucoma, and Diabetic Retinopathy. Each category comprises approximately 1000 retinal images, which are acquired from standard sourced like IDRiD, Ocular Recognition, and HRF, respectively.

3.2 Generative adversarial network based adversarial attack model

After the acquisition of medical images, these data are fed as input to the proposed Wasserstein GAN model to generate synthetic images from the actual medical images. The GAN model is widely used in data augmentation to address the data imbalance issue, which helps to enhance detection or classification performance. Here, the GAN model is used to generate various types of AA during adversarial training of the model. There are two neural networks in the GAN architecture: Generator

and Discriminator, where the generator is used to create AA and the discriminator is used to analyze the actual attack and the generated adversarial attack. A mathematical representation of conventional GAN is represented in Eq. (1):

$$min_{\theta G}\, max_{\theta D} V(G,D) = min_G\, max_D min_{\theta G}\, E_{x\sim Pdata}[\log D(x)] + E_{z\sim Pz}\left[\log\left(1-D\left(G(z)\right)\right)\right] \quad (1)$$

Where $G(z)$ denotes adversarial attack; (x) represents the actual attack; z indicates the input vector; D stands for the classification probability. However, the traditional GAN has a limitation, which is non-convergence, meaning that the generator network is unable to learn an optimal adversarial distribution.

This non-convergence affects adversarial images generated that do not resemble medical images closely and makes them easy to detect by adversarial defense mechanisms. Thus, an improved GAN model, namely Wasserstein GAN (WGAN), is used in this research to improve convergence and enhance adversarial attack classification. The major difference between conventional GAN and WGAN is distinguished by estimating the difference between two distributions. The WGAN utilizes a Wasserstein distance instead of Jensen Shannon Divergence (JSD), making the network more stable than the conventional GAN. The mathematical representation of the proposed WGAN is given in Eq. (2):

$$W(p_r,p_g) = \inf_{\gamma\sim\pi(p_r,p_g)} E_{(x,y)\sim\gamma}\|x-y\| \quad (2)$$

Where p_r, p_g represents the distributions; $\pi(p_r,p_g)$ denotes the lowest term;

3.3 Adversarial attacks (AA)

The adversarial attack is developed to manipulate input medical image, which is imperceptible to human vision, but it affects the classification model to provide misclassification results. Various attacks are considered in this research, such as FGSM, PGD, and Noise perturbation attacks, respectively. These attacks are described as follows:

3.3.1. FGSM attacks

The FGSM attack is an effective attack that is widely considered in the adversarial training of medical images that is mentioned in previous research works. This attack is based on gradient computation in a single step to create perturbations that are controlled within a particular form in the medical images. The primary objective of this FGSM attack is to reduce efficiency of perturbation calculations of DL models rather than focusing on high false rates. The nature of an FGSM attack is that it perturbs the input image by inserting noise in the direction of the input gradient, which is given in Eq. (3):

$$x_{adv} = x + \in .\, sign\left(\nabla_x \zeta(h(x),y)\right) \quad (3)$$

Where y represents the true label of the input image; $h(x)$ indicates the output of the neural network model; $sign\,(.)$ indicates computation of the sign of the gradient; ∇_x denotes the gradient for input; $\zeta\,(.)$ stands for the loss function.

3.3.2. PGD attacks

The PGD is the strongest first-order attack that generates perturbation. Instead of using a single step, this attack updates multiple iterations of the gradients in the FGSM attack. This PGD attack is mathematically expressed in Eq. (4):

$$x_{adv}^{t} = \prod_{\epsilon}\left(x_{adv}^{t-1} + \alpha.\, sign\left(\nabla_x \zeta(h(x_{adv}^{t}),y)\right)\right) \quad (4)$$

Where x^t represents the adversarial examples at the t^{th} step; $t-1$ denotes AA at the previous step; h indicates the model prediction function.

3.3.3. Noise perturbation attack

Noise perturbation is one of the AAs that introduces noise randomly into the input medical images, which manipulates prediction results generated by the classification model. The attacker aims to interrupt the DL-based classification model's decision-making process, which results in imprecise predictions. Here, the attacker alters medical images by inserting a carefully crafted noise or a random noise. Since the inserted noise in the images is visible to humans, it can affect the model's output significantly. The goal of this attack is to control the decision of the model that makes it more vulnerable to misclassifications or producing unreliable results.

The generated synthetic medical images with AA are fed to the preprocessing phase to enhance image quality.

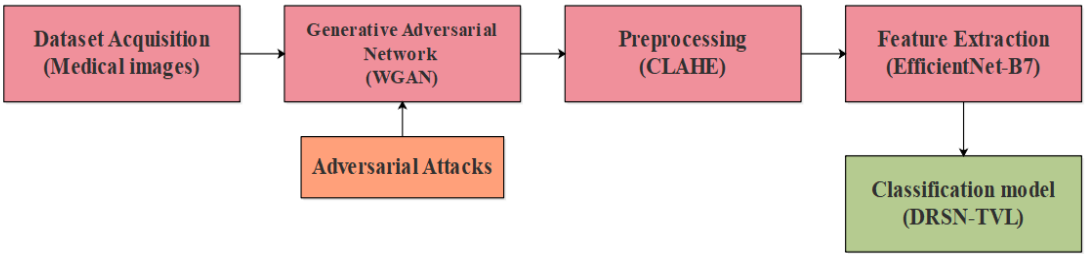

Figure. 1 Block diagram of proposed identification and classification of AA of the medical images.

3.4 Preprocessing

The generated adversarial images are passed to the preprocessing phase to improve image quality for further identification/ detection processes. In this research, a Contrast Limited Adjustment Histogram Equalization (CLAHE) is utilized to improve contrast in certain portions to identify the disease efficiently. CLAHE is one of the most widely used preprocessing techniques to enhance image contrast, which equalizes the image histogram within small areas. Here, a weighted Gaussian blur in the CLAHE technique which utilize a weight to control the amount of blur at every pixel in the image. These weights are typically obtained by the pixel's distance from the blur's center of the image to estimate the Gaussian function in two dimensions (x, y), which is mathematically represented in Eq. (5):

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

Where x and y denote distances along the horizontal and vertical axis from origin; σ represents the standard deviation of Gaussian distribution. This CLAHE technique improves image contrast in smaller regions, which makes the model more prone to misclassifications when adversarial noise is introduced. These pre-processed images are then forwarded as input to the improved GAN model to generate fake images.

3.5 Feature extraction

After training the AA into the medical images, these images are fed as input to the pre-trained model-based feature extraction process. In this research, EfficientNet-B7 is utilized for efficient feature extraction from the actual medical images and the attack-incorporated medical images. Generally, the EfficientNet-B7 model is distributed into seven various blocks, where each block has several number of Mobile inverted Bottleneck Convolution (MBConv), and each MBConv block has various sizes of filters, channels, and strides. Hence, this pre-trained model performs feature extraction by utilizing compound scaling method, which involves depth, width, and resolution. These scales are adjusted to achieve a balance between model accuracy, size, and computational efficiency to ensure that model maintains a well-proportioned and optimized structure. The mathematical representation of the EfficientNet-B7 model is given in Eqs. (6) – (8):

$$Depth\ (d) = \alpha^{\varphi} \tag{6}$$

$$Width\ (w) = \beta^{\varphi} \tag{7}$$

$$Resolution\ (r) = \gamma^{\varphi} \tag{8}$$

Where α, β and γ denotes constants for depth, width and resolution; φ (Phi) represents the compound coefficient.

The EfficientNet-B7 has a highly expressive feature extraction capability due to its compound scaling approach based on width, depth, and resolution. This permits the AA to exploit richer features and ensures that perturbations impact critical diagnostic patterns effectively. Especially, adversarial perturbations on chest x-ray images using the EfficientNet-B7 model exploit fine-grained lung abnormalities effectively compared to other shallower networks. These extracted features are further fed as input to the proposed classification model.

3.6 Proposed DRSN classification model

The extracted features are fed as input to the proposed DRSN-TVL-based classification model to effectively classify the medical images. In this research, a shrinkage network with a ResNet model as the backbone network, along with a Total variation loss, is incorporated into the network to enhance the classification of AA-based medical images precisely. A ResNet is utilized in this research, which stacks many distinct residual building units that include Convolutional layers (Convs), Batch Normalization layers (BNs), and the Rectified Linear Unit (ReLU)-based activation function. This ResNet architecture

contains skip connection between various layers to mitigate the vanishing gradient issue and achieve superior performance to traditional convolutional neural network architectures. In this DRSN model, a soft thresholding function is extensively utilized as a primary step in several signal-denoising techniques, especially in adversarial training, to learn about noise perturbations from the extracted features from Efficient-B7 and identify the AA efficiently. The soft thresholding of DRSN is mathematically expressed in Eq. (9):

$$y = \begin{cases} x - \tau & x > \tau \\ 0 & -\tau \le x \le \tau \\ x + \tau & x < -\tau \end{cases} \quad (9)$$

Where x and y denote input and output features in DRSN; τ represents threshold, which is a positive parameter. To preserve negative features, it sets near zero features to zeros. Accordingly, the partial derivative of this soft thresholding function is mathematically expressed in Eq. (10):

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & x > \tau \\ 0 & -\tau \le x \le \tau \\ -1 & x < -\tau \end{cases} \quad (10)$$

This soft thresholding function is integrated into the residual shrinkage block unit to remove noise-related features from the extracted features, which lead to misclassification results. The residual shrinkage block unit is incorporated in the convolutional layer to acquire the most relevant features from the extracted features of AA-based medical images. Selecting relevant features, which has significant information about the AA helps to identify and classify the attacks in the medical images efficiently.

3.6.1. Total variation loss

The total variation loss is used to make the actual medical images and adversarial attack-based images smoother and reduce the influence of noise during the adversarial training process, which is represented in Eq. (11):

$$L_{TV} = \sum_{m,n} \left(\left(I_{m,n+1} - I_{m,n} \right)^2 + \left(I_{m,1} - I_{m,n} \right)^2 \right) \quad (11)$$

Where m and n denote the position of the coordinate of the pixel in medical images I.

This research utilized several techniques, like CLAHE, to enhance medial image contrast, and WGAN to generate fake medical images to integrate AAs such as FGSM, PGD, and noise perturbation. After that, EfficientNet-B7 is used for feature extraction, which has the advantage of being more computationally efficient than deeper architectures like ResNet-152 or Inception-v4 despite its large capacity. This enables faster adversarial training, reducing the time required to generate high-quality adversarial examples. The proposed DRSN model utilizes soft-thresholding layers to remove irrelevant noise while preserving important medical features. Also, the TVL function reduces unnecessary variations in pixel intensities, ensuring perturbations remain smooth and realistic. This loss function prevents adversarial perturbations from introducing unnatural high-frequency noise, making them harder to detect by adversarial defence mechanisms.

4. Results and discussion

Experimental results of the proposed DRSN-TVL method, utilized for adversarial training for AA in medical images using chest X-ray datasets, eye disease datasets, and 3D- CT Lung nodule datasets, are depicted in this section. The proposed model is simulated on the Python 3.9 software tool with a system configuration of Windows 10, 16 GB RAM, and an i5 processor. The qualitative and quantitative analysis of GAN, the feature extraction model, and the proposed classification model with their state-of-the-art approaches is analyzed in section 4.1. The comparative analysis of the proposed classification method is evaluated with existing classification methods used for adversarial training in section 4.2. based on different performance metrics.

Table 1. Represents the parameter settings of the proposed model used to identify the AA in medical images.

The performance metrics like Accuracy, Precision, Sensitivity, and F1-score are used in this research to estimate the effectiveness of posed DRSN-TVL method in adversarial training. Mathematical representation of the performance metrics is expressed in Eqs. (12) to (15):

Table 1. parameter settings of proposed DRSN-TVL model

Parameters	Value
Learning rate	0.00001
Activation function	Softmax
Dropout	0.3
Optimizer	Adam
Loss function	Total Variation Loss

Table 2. Performance of WGAN based on Chest X-Ray dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
GAN	Chest X-Ray	96.25	94.15	93.10	93.62
DCGAN		97.35	95.14	92.95	94.03
WGAN		98.45	98.13	97.89	98.00

Table 3. Performance of WGAN based on Fundoscopy dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
GAN	Fundoscopy	96.49	96.48	96.47	96.47
DCGAN		97.61	97.60	97.62	97.61
WGAN		98.73	98.26	97.92	98.08

Table 4. Performance of EfficientNet-B7 based on Chest X-Ray dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
ResNet	Chest X-Ray	93.33	91.88	91.50	91.88
DenseNet		95.38	93.23	92.40	92.81
VGG		95.94	95.93	95.91	95.92
InceptionNet		96.81	96.80	96.78	96.79
MobileNet		97.42	96.10	95.07	95.58
EfficientNet-B7		98.45	98.13	97.89	98.00

Table 5. Performance of EfficientNet-B7 based on Fundoscopy dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
ResNet	Fundoscopy	94.35	94.34	94.33	94.33
DenseNet		94.79	94.78	94.77	94.77
VGG		95.84	95.82	95.83	95.82
InceptionNet		96.93	96.92	96.91	96.91
MobileNet		97.68	97.66	96.67	97.16
EfficientNet-B7		98.73	98.26	97.92	98.08

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \quad (13)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (14)$$

$$F1-Score = 2 \times \frac{Precision \times Sensitivity}{Precision+ensitivity} \quad (15)$$

Where TP and FP denote True Positive and False Positive; TN and FN represent True Negative and False Negative.

4.1 Quantitative and Qualitative Analysis

Performance analysis of DRSN-TVL method, which is utilized for adversarial attack identification in medical images using two benchmark datasets, is depicted in this research. The performance evaluation of DRSN-TVL method utilizes WGAN for generating fake images for adversarial training and classifying the medical images. Table 2 and Table 3 represent the performance of WGAN with various types of GAN models.

Performance evaluation of e proposed model utilizes a deep learning-based pre-trained model for feature extraction from the pre-processed medical images. Tables 4 and 5 represents the performance of the EfficientNet-B7-based feature extraction model. The state-of-the-art pretrained models are Residual Network (ResNet), DenseNet, MobileNet, and conventional neural network.

Table 6. Performance of proposed DRSN-TVL based classification method in Chest X-ray dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
CNN	Chest X-Ray	94.31	93.54	92.11	92.81
DNN		95.22	95.01	93.41	94.20
Transformer		97.62	96.60	95.20	95.89
Proposed DRSN-TVL		98.45	98.13	97.89	98.00

Table 7. Performance of proposed DRSN-TVL based classification method in Fundoscopy dataset

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
CNN	Fundoscopy	95.72	95.71	95.70	95.70
DNN		96.61	96.60	96.62	96.61
Transformer		97.91	97.90	97.89	97.89
Proposed DRSN-TVL		98.73	98.26	97.92	98.08

Table 8. Performance of proposed DRSN classification method based on with and without TVL function

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
DRSN (without TVL)	Chest X-Ray	96.72	96.10	95.80	95.94
DRSN-TVL		98.45	98.13	97.89	98.00
DRSN (without TVL)	Fundoscopy	96.90	96.71	96.60	96.65
DRSN-TVL		98.73	98.26	97.92	98.08

Table 9. Performance of proposed DRSN-TVL based classification method based on computational complexity

Dataset	Methods	Model Complexity	
		Params (M)	Flops (G)
Chest X-Ray	CNN	11.0	1.7
	DNN	10.9	1.5
	Transformer	10.8	1.4
	Proposed DRSN-TVL	10.6	1.2
Fundoscopy	CNN	10.8	1.5
	DNN	10.9	1.4
	Transformer	10.7	1.3
	Proposed DRSN-TVL	10.6	1.2

Performance evaluation of the proposed model utilized for identification and classification of medical images after adversarial training. Tables 6 and 7 represent performance of DRSN-TVL method-based classification model. The existing classification models are CNN, Deep neural Network (DNN), and Transformer, which are utilized for evaluation in this section. The proposed DRSN model utilizes soft-thresholding layers to remove irrelevant noise while preserving important medical features. Also, the TVL function reduces unnecessary variations in pixel intensities, ensuring perturbations remain smooth and realistic. Thus, the proposed DRSN-TVL achieves better results in medial image classification with AA efficiently.

Performance evaluation of the proposed DRSN-TVL method based on with and without TVL loss function using Chest X-Ray and Fundoscopy datasets. Table 8 illustrates the performance DRSN model in terms of presence and absence of loss function. The proposed TVL function effectively reduces high-frequency noise and prevents the DRSN model from learning false adversarial patterns, thereby improving generalization.

Performance analysis of DRSN-TVL method is evaluated based on computational complexity based on parameters and flops. Table 9 illustrates the effectiveness of DRSN-TVL model with other classifier models in terms of computation complexity. The proposed method attains less computational complexity for the identification of adversarial attacks in medical images.

Performance analysis of the proposed DRSN-TVL method based on various stronger attacks for

Chest X-Ray and Fundoscopy datasets is represented in Table 10. The effectiveness of DRSN-TVL method is analysed by evaluating the model with various stronger adversarial attacks such as Carlini & Wagner (CW), DeepFool and Autoattack (ensemble of robust, parameter-free attacks). From the table 10, the results show that the DRSN-TVL model maintains strong performance even against stronger and more adaptive adversarial attacks that ensure its robustness across the evaluated attacks.

Performance analysis of the proposed DRSN-TVL method based on for standard adversarial pipelines using Chest X-Ray and Fundoscopy datasets is illustrated in Table 11. The effectiveness of DRSN-TVL method is analysed by evaluating the model with standard adversarial attacks such as FGSM, PGD, CW, DeepFool, Boundary attack and Autoattack respectively. From the table 10, the results show that the DRSN-TVL model achieved better results in identifying and classifying the standard adversarial attacks effectively which generated independently. Moreover, these results ensure that the proposed DRSN-TVL model robustness is not depends on specific WGAN-perturbation characteristics.

4.2 Comparative analysis

Comparative study of proposed DRSN-TVL method utilizes the classification of medical images after adversarial training. Table 12 represents the performance of DRSN-TVL -based classification model for the chest X-ray dataset. Table 13 ilustrates performance of DRSN-TVL method-based classification model for fundoscopy dataset. Table 14 represents the performance analysis of the proposed method-based classification model for the Dermatology dataset. Table 15 represents the effectiveness of DRSN-TVL method-based classification model for Eye dataset and 3D-Lung nodule dataset.

Table 10. Performance of proposed DRSN-TVL based classification method based on various adversarial attacks

Dataset	**Attacks**	**Metrices**	
		Accuracy (%)	**F1-score (%)**
Chest X-Ray	CW	95.28	95.10
	DeepFool	94.70	94.45
	AutoAttack	93.85	93.60
Fundoscopy	CW	96.32	96.10
	DeepFool	95.85	95.67
	AutoAttack	94.93	94.71

Table 11. Performance of proposed DRSN-TVL based classification method for standard adversarial pipelines

Dataset	**Attacks**	**Metrices**	
		Accuracy (%)	**F1-score (%)**
Chest X-Ray	FGSM (CleverHans)	96.87	96.65
	PGD (Foolbox)	95.72	95.50
	DeepFool (Foolbox)	94.89	94.63
	CW (Foolbox)	95.10	94.88
	Boundary attack (Foolbox)	93.52	93.23
	AutoAttack	93.85	93.60
Fundoscopy	FGSM (CleverHans)	97.15	96.92
	PGD (Foolbox)	96.23	96.05
	DeepFool (Foolbox)	95.72	95.48
	CW (Foolbox)	95.90	95.70
	Boundary attack (Foolbox)	94.85	94.63
	AutoAttack	94.26	94.02

Table 12. Comparative analysis of proposed DRSN-TVL classification method

Methods	**Dataset**	**Metrices**			
		Accuracy (%)	**Precision (%)**	**Sensitivity (%)**	**F1-score (%)**
ResNet-152 V2 [16]	Chest X-Ray	81.23	N/A	N/A	N/A
DenseNet [17]		96.85	95.86	96.16	96.00
ResNet-18 [20]		86.66	N/A	N/A	N/A
Proposed DRSN-TVL		98.45	98.13	97.89	98.00

Table 13. Comparative analysis of proposed DRSN-TVL classification method

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
ResNet-18 [20]	Fundoscopy	87.73	N/A	N/A	N/A
Proposed DRSN-TVL		98.73	98.26	97.92	98.08

Table 14. Comparative analysis of proposed DRSN-TVL with existing methods

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
ResNet-18 [20]	Dermatology	66.85	N/A	N/A	N/A
Proposed DRSN-TVL		96.42	96.41	96.40	96.40

Table 15. Comparative analysis of proposed DRSN-TVL with existing methods

Methods	Dataset	Metrices			
		Accuracy (%)	Precision (%)	Sensitivity (%)	F1-score (%)
EFF model [21]	Eye dataset	96.8	93.9	95.1	94.5
Proposed DRSN-TVL		97.5	97.6	97.4	97.4
EFF model [21]	3D – CT lung nodule	92	90	93	91
Proposed DRSN-TVL		94.7	94.6	94.4	94.5

The existing classification models are ResNET-152 v2 [16], DenseNet [17], ResNet-18 [20], and EFF model [21], which are utilized for comparative analysis in this section.

The comparative analysis of the proposed method with the EFF model [21] is performed based on Scenario 3 (S3) only in [21], which describes that the model performed on adversarial attack inserted eye dataset and 3D–CT lung nodule dataset. The performance of the proposed DRSN-TVL achieves better results than the ensemble feature fusion model because the shrinkage network only considers the most relevant features to identify the attacks in medical images efficiently.

4.3 Discussion

The proposed DRSN-TVL model achieves better results in medical image classification by identifying the AA efficiently. The proposed DRSN model can shrink redundant features and force the network to rely more on vulnerable feature representations, which make the model enhance the classification performance. The TVL function prevents adversarial perturbations from unnatural high-frequency noise, which makes the model difficult to detect by adversarial defense mechanisms. Thus, the proposed method achieves an accuracy of 98.45% in the Chest X-Ray dataset, 98.73% in the Fundoscopy dataset, 97.5% in the Eye disease dataset, and 94.7% in the 3D-CT Lung nodule dataset, which is better than the existing deep neural network-based classification model. The exiting models like ResNet-152 V2 [16], DenseNet [17], ResNet-18 [20], and EFF model [21] have limitations such as difficulty in identifying the AA due to redundant features and high noise perturbations. By addressing these limitations efficiently, the proposed DRSN-TVL model achieves high accuracy classification of medical images, which effectively identifies the AA and noise perturbations.

5. Conclusion

The DRSN-TVL method is proposed to identify the AA and classify the medical images accurately. The proposed DRSN model shrinks redundant features and removes irrelevant noise by soft thresholding that preserves important medical image features. Moreover, the proposed TVL function, which is incorporated with the DRSN model, minimizes unnecessary variations in pixel intensities that ensure the perturbations in the image remain smooth and realistic. Initially, the medical images are acquired from Chest X-ray, Fundoscopy, and Eye disease datasets. Next, CLAHE is used to enhance the image contrast, and then WGAN is used for generating fake images to integrate the AA. After that, an EfficientNet-B7 model is utilized for extracting the features of AA, and based on the extracted features, the proposed DRSN-TVL model identifies the attacks effectively. The experimental results of the proposed DRSN-TVL method achieve an accuracy of 98.45% in the chest X-ray dataset, which is higher than the existing method, ReseNet-153 V2. In future, advanced DL-based approaches will be

used for improving the identification of AA in the medical images effectively.

Notation list

Notation	Description
$G(z)$	Adversarial attack
(x)	Actual attack
z	Input vector
D	Classification probability
p_r, p_g	Distributions
$\pi(p_r, p_g)$	Lowest term
y	True label of the input image
$h(x)$	Output of the neural network model
$ign(.)$	Computation of sign of gradient
∇_x	Gradient for input
$\zeta(.)$	Loss function
x^t	Adversarial examples at t^{th} step
$t-1$	Adversarial attacks at the previous step
h	Model prediction function
x and y	Distance along horizontal and vertical axis from origin
σ	Standard deviation of Gaussian distribution
α, β and γ	Constants for depth, width and resolution
φ (Phi)	Compound coefficient
x and y	Input and output feature in DRSN
τ	Threshold which is a positive parameter
m and n	Position of the coordinate of the pixel in medical images I

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, AG and NJ; methodology, US; software, AG; validation, NJ, US, and AG; formal analysis, NJ; investigation, US; resources, AG; data curation, NJ; writing—original draft preparation, US; writing—review and editing, NJ; visualization, AG; supervision, AG; project administration, AG; funding acquisition, US.

References

[1] B.U.H. Sheikh and A. Zafar, "Removing adversarial noise in x-ray images via total variation minimization and patch-based regularization for robust deep learning-based diagnosis", *Journal of Imaging Informatics in Medicine*, Vol. 37, pp. 3282–3303, 2024.

[2] S. Pal, S. Rahman, M. Beheshti, A. Habib, Z. Jadidi, and C. Karmakar, "The Impact of Simultaneous Adversarial Attacks on Robustness of Medical Image Analysis", *IEEE Access*, Vol. 12, pp. 66478-66494, 2024.

[3] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications", *Artificial Intelligence Review*, Vol. 58, No. 1, pp. 1-107, 2025.

[4] A.S. Neha, V. Chaturvedi, and M. Shafique, "FRNet: A Feature-Rich CNN Architecture to Defend against Adversarial Attacks", *IEEE Access*, Vol. 12, pp. 26943-26956, 2024.

[5] K.V. Priya and J. Dinesh Peter, "Enhanced defensive model using CNN against adversarial attacks for medical education through human computer interaction", *International Journal of Human–Computer Interaction*, Vol. 41, No. 3, pp. 1729–1741, 2023.

[6] M. Alkhowaiter, H. Kholidy, M.A. Alyami, A. Alghamdi, and C. Zou, "Adversarial-aware deep learning system based on a secondary classical machine learning verification approach", *Sensors*, Vol. 23, No. 14, p. 6287, 2023.

[7] Y. Kim, J. Jung, H. Kim, H. So, Y. Ko, A. Shrivastava, K. Lee, and U. Hwang, "Adversarial Defense on Harmony: Reverse Attack for Robust Models Against Adversarial Attacks", *IEEE Access*, Vol. 12, pp. 176485-176497, 2024.

[8] S.H. Zhong, S. Zhao, Z. Xiao, Z. Zhang, and Y. Liu, "Attention-guided universal adversarial perturbations for EEG-based brain–computer interfaces", *Expert Systems with Applications*, Vol. 268, p. 126362, 2025.

[9] M. Xu, T. Zhang, and D. Zhang, "Medrdf: a robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack", *IEEE Transactions on Medical Imaging*, Vol. 41, No. 8, pp. 2130-2143, 2022.

[10] A.A. Abd El-Aziz, R.A. El-Khoribi, and N.E. Khalifa, "RDMAA: Robust Defense Model against Adversarial Attacks in Deep Learning for Cancer Diagnosis", *International Journal of Computing and Digital Systems*, Vol. 15, No. 1, pp. 1273-1287, 2024.

[11] E. Kanca, S. Ayas, E. Baykal Kablan, and M. Ekinci, "Evaluating and enhancing the robustness of vision transformers against adversarial attacks in medical imaging", *Medical & Biological Engineering & Computing*, Vol. 63, pp. 673-690, 2024.

[12] L. Alzubaidi, A.D. Khamael, H.A.H. Obeed, A. Saihood, M.A. Fadhel, S.A. Jebur, Y. Chen, A.S. Albahri, J. Santamaría, A. Gupta, and Y. Gu,

"MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging", *Intelligent Systems With Applications*, Vol. 22, p. 200355, 2024.

[13] A. Hanif, F. Shamshad, M. Awais, M. Naseer, F.S. Khan, K. Nandakumar, S. Khan, and R.M. Anwer, "Baple: Backdoor attacks on medical foundational models using prompt learning", In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 443-453, 2024.

[14] H. Ding, N. Huang, Y. Wu, and X. Cui, "LEGAN: Addressing Intra-class Imbalance in GAN-based Medical Image Augmentation for Improved Imbalanced Data Classification", *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, p. 2517914, 2024.

[15] E. Kanca, S. Ayas, E.B. Kablan, and M. Ekinci, "Evaluating and enhancing the robustness of vision transformers against adversarial attacks in medical imaging", *Medical & biological engineering & computing*, Vol. 63, pp. 673-690, 2024.

[16] D. Vasan, and M. Hammoudeh, "Enhancing resilience against adversarial attacks in medical imaging using advanced feature transformation training", *Current Opinion in Biomedical Engineering*, Vol. 32, p. 100561, 2024.

[17] B.U.H. Sheikh, "Mitigating adversarial threats in deep CT image diagnosis models via a dual-stage inference-time defense", *Applied Soft Computing*, Vol. 163, p. 111909, 2024.

[18] P. Laykaviriyakul, and E. Phaisangittisagul, "Collaborative Defense-GAN for protecting adversarial attacks on classification system", *Expert Systems with Applications*, Vol. 214, p. 118957, 2023.

[19] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, and A. Ayatollahi, "MedViT: a robust vision transformer for generalized medical image classification", *Computers in Biology and Medicine*, Vol. 157, p. 106791, 2023.

[20] Y. Dai, Y. Qian, F. Lu, B. Wang, Z. Gu, W. Wang, J. Wan, and Y. Zhang, "Improving adversarial robustness of medical imaging systems via adding global attention noise", *Computers in Biology and Medicine*, Vol. 164, p. 107251, 2023.

[21] L. Alzubaidi, A.D. Khamael, H.A.H. Obeed, A. Saihood, M.A. Fadhel, S.A. Jebur, Y. Chen, A.S. Albahri, J. Santamaría, A. Gupta, and Y. Gu, "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging", *Intelligent Systems With Applications*, Vol. *22*, p. 200355, 2024.

[22] ChestXray dataset: https://www.kaggle.com/datasets/nih-chest-xrays/data (Accessed on February 2025)

[23] Fundoscopy dataset (Kaggle Aptos 19 blindness dataset): https://www.kaggle.com/datasets/mariaherrerot/aptos2019 (Accessed on February 2025)

[24] Al-Shabi, K. Shak, and M. Tan, "ProCAN: Progressive growing channel attentive non-local network for lung nodule classification", *Pattern Recognition*, Vol. *122*, p. 108309, 2022.

[25] Eye datset: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification (Accessed on February 2025)